DOCUMENT RESUME

ED 413 349 TM 027 657

AUTHOR	Land, Robert
TITLE	Moving Up to Complex Assessment Systems. Proceedings from the CRESST Conference (Los Angeles, CA, September 5-6, 1996).
INSTITUTION	National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	1997-00-00
NOTE	25p.
CONTRACT	R305B60002
AVAILABLE FROM	UCLA Center for the Study of Evaluation, 10920 Wilshire Boulevard, Suite 900, Los Angeles, CA 90024-6511.
PUB TYPE	Collected Works - Proceedings (021) -- Collected Works - Serials (022) -- Reports - Evaluative (142)
JOURNAL CIT	Evaluation Comment; v7 n1 p1-22 Sum 1997
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Conferences; *Educational Assessment; Educational Research; Educational Technology; Elementary Secondary Education; *Evaluation Utilization; Models; Reliability; Research and Development; *Standards; *Test Use; Validity
IDENTIFIERS	*Center for Research on Eval Standards Stu Test CA

ABSTRACT

Assessment systems to measure high educational standards emerged as the major theme at the 1996 conference of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), "Moving Up to Complex Assessment." This feature article, providing a summation of the proceedings of the conference, reports that the approximately 250 educators and community leaders were in general agreement that challenging standards are the key to the improvement of American education. In opening remarks, the codirectors of CRESST, Eva L. Baker and Robert L. Linn explained the conceptual model that is guiding CRESST assessment research and development in the next 5 years. This model focuses on the utility of assessment systems for various purposes and establishes long-range goals for the Center's research. The CRESST model highlights three qualities that are essential to the productive use of assessment: validity, fairness, and credibility. Conference presentations centered on broad areas related to these qualities: (1) developing valid, fair, and credible assessments; (2) enhancing the utility of assessments; and (3) exploring the role that technology can play in creating new possibilities in developing and using assessment systems. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

ED 413 349

EVALUATION COMMENT

A PUBLICATION OF UCLA'S CENTER FOR THE STUDY OF EVALUATION AND GRADUATE SCHOOL OF EDUCATION & INFORMATION STUDIES

NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS, AND STUDENT TESTING

Summer 1997, Vol. 7, No. 1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

MOVING UP TO COMPLEX ASSESSMENT SYSTEMS

PROCEEDINGS FROM THE 1996 CRESST CONFERENCE

ROBERT LAND UCLA/CRESST

ASSESSMENT systems to measure high educational standards emerged as the major theme at this year's CRESST conference, Moving Up to Complex Assessment Systems, September 5-6, 1996, at UCLA's Sunset Village Conference Center. The broad appeal of the agenda was reflected by the diversity of the conference participants. Approximately 250 researchers, community leaders, teachers, principals, school board members, state and federal education officials, and representatives of private and commercial interests attended two full days of presentations and assessment forums.

Opinion was strong from many conference presenters that challenging standards were the backbone to the improvement of American education.

MAILING LIST UPDATE

We will be updating the CRESST mailing list during the next few months. Postcards will be mailed to everyone who currently receives free copies of *CRESST Line* and *Evaluation Comment*. To remain on our list, please return the postcard promptly.

Just a reminder that you or your associates may register to be placed on our publications mailing list at any time through our Web site, www.cse.ucla.edu.

Thank you for your cooperation.

TM027657

BEST COPY AVAILABLE

"I'd rather set the [standards] bar high," said Sidney Thompson, superintendent of the Los Angeles Unified School District, "and have us look at how we're going to help the student get over that bar, than set the bar low and know that when she or he got over it, it didn't mean a darn thing."

Thompson noted that the Los Angeles Unified School District has joined all 50 states and many large school districts in setting standards for what children should know and be able to do across multiple grade levels and topics. The District and CRESST are working together to develop a new standards-based assessment system comprised of a commercial standardized test, performance-based assessments based on CRESST instructional models, and classroom tests to improve instruction, learning, and student performance.

"Our current emphasis on high, challenging standards for all students," said CRESST Co-director Eva Baker in her conference presentation, "can be traced to the 1989 Governors' Education Summit and was reinforced in the 1994 Goals 2000 legislation and the recent Improving America's Schools Act[1] which reauthorized federal Title I programs."

"By 1997-98," explained Baker, "Title I schools must have in place challenging content and performance standards in at least reading and mathematics, followed by high-quality assessments of those standards in the 2000-2001 school year. The assessments must involve multiple approaches and measure complex thinking skills and understanding of rigorous content. Performance assessments will be part of the system but pose some unique challenges in terms of the

[1]Improving America's Schools Act of 1994, Conference Report 103-761. Regarding Public Law 103-382, signed October 20, 1994, (pp. 6-33). Washington, DC: House of Representatives.

UCLA's
Center for the
Study of Evaluation
&
The National Center
for Research on Evaluation,
Standards, and Student Testing

Eva L. Baker, Co-Director
Robert L. Linn, Co-Director
Joan L. Herman, Associate Director
Pamela Aschbacher, Assistant Director
Ronald Dietel, CRESST Director of Communcations
Katharine Fry, Editorial Assistant
Mary Wilby, Design Layout

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this publication do not reflect the position or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

time, costs and technical quality required to develop accurate measures of individual accomplishment."

"One of our biggest challenges is to learn how performance assessments can work with traditional assessments," noted CRESST Co-director Robert Linn in his opening conference remarks, "and how these assessments fit into the larger education reform picture, from the classroom to the national level. We've moved from a focus on single instruments to a system perspective."

In other opening remarks, both Baker and Linn explained the new conceptual model that is guiding the CRESST assessment research and development efforts for the next five years. The CRESST model (Figure 1) focuses on the utility of assessment systems for various purposes and audiences and establishes three important, long-range social goals for the Center's research:

- providing new knowledge and understanding about educational quality;
- contributing to educational improvement in policy, accountability, and teaching and learning; and
- encouraging productive public engagement in education.

The CRESST model highlights three qualities that are essential to the productive use of assessment, namely, validity, fairness and credibility, and focuses the CRESST research agenda on understanding the relationships among and between these qualities and effective assessment systems.

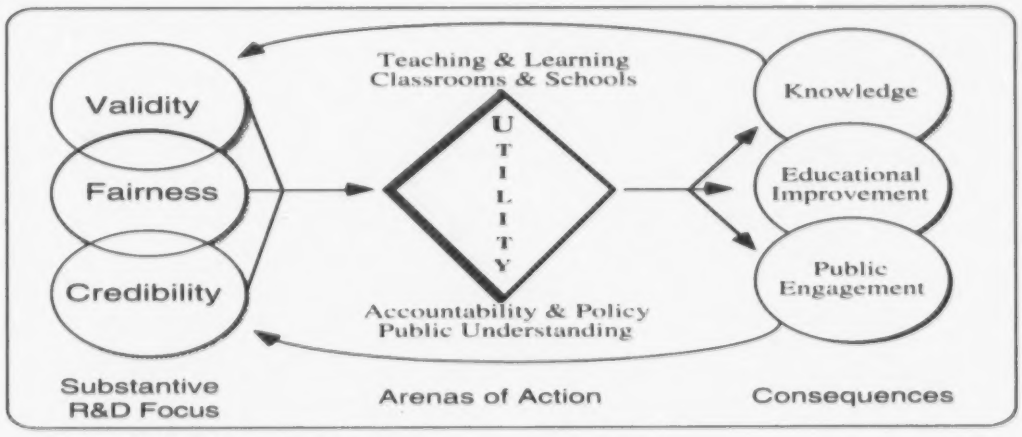

Figure 1. CRESST Conceptual Model

Conference presentations focused on three broad areas related to the CRESST conceptual model: developing valid, fair, and credible assessments; enhancing the utility of assessments; and finally, the role that technology can play in creating new possibilities in both developing and utilizing assessment systems.

DEVELOPING VALID, FAIR, AND CREDIBLE ASSESSMENTS

In the past, researchers, educators, and policy makers seemed satisfied if assessments met relatively narrow criteria of technical quality. But as the purposes of assessment have grown and the demand for more inclusive and informative tests has increased, both test developers and test users have recognized that narrow technical criteria are not enough. Echoing themes from the CRESST model, conference presenters took a comprehensive view of what is required for good assessment: an expanded view of validity, including attention to intended purposes and consequences; heightened concerns for inclusion of and fairness to all students; and recognition of the importance of public credibility.

Please note that beginning with this issue of *Evaluation Comment* we shall resume the use of volume and issue numbers.

Validity

Today's assessment systems are intended to serve multiple audiences at the federal, state, local, classroom, and student levels and likewise are intended to serve a range of purposes, from communicating standards and promoting accountability, to contributing to school improvement, informing teaching and learning, and improving student performance. These demands bring new complexity to assuring that assessment systems provide accurate information for decision making. Conference participants particularly highlighted three areas warranting sustained effort: alignment, the measurement of progress, and linking the results from multiple measures.

> "...assessment is both a very central part of reform and the index for judging the success of that reform."

Alignment

That assessment is to be aligned with rigorous standards for student achievement is a defining feature of today's assessments. As Ed Reidy, deputy commissioner of the Kentucky Department of Education expressed it, "Assessment is both

a very central part of reform and the index for judging the success of that reform." Assessment is intended to stimulate reform by communicating these standards, holding educators and students accountable for achieving them, and to provide an accurate measure of students' performance on the standards.

"It seems simple—adopt standards and make assessments that are aligned with them—but there is a lot more involved," noted Robert Linn.

What does such alignment really mean? How do states, districts, and schools know whether their assessments are aligned? "How the major elements of an education system work together to guide the process of helping students achieve higher levels of mathematical and scientific understanding," said Norman Webb, Wisconsin Center for Educational Research, "goes beyond a simple content analysis."

Citing results of a study by the Council of Chief State School Officers (CCSSO), Webb reported that few states have addressed these questions with much rigor and even fewer have examined broader alignment issues. Contributing to the complex alignment picture is that states lack a formal and systematic process to develop assessments directly based on their standards, and instead have developed assessments prior to or at the same time as their standards.

Based on his review of current practices and relevant literature, Webb presented five categories of criteria that states and local districts can use to evaluate alignment.

These categories include:

- pedagogical implications;
- equity and fairness;
- articulation across grades and ages;
- system applicability; and
- content focus.

Content focus includes topic coverage, depth and range of student knowledge, and balance of representation.

Several presenters reported that even when assessments are developed directly from standards, alignment can be complicated. David Wiley, technical director of the New Standards Project, summed up the general problem by stating that standards—even performance standards anchored in students' work—are not specified well enough for purposes of test development. They do not adequately guide the concrete decisions that need to be made on what is to be measured, how it is to be measured, and what specific tasks and criteria will be used.

To develop the New Standards mathematics assessments, Wiley found it was necessary to create an "infrastructure that would provide another level of construct definition." The New Standards' seven measurable mathematics standards, for example, were clustered under three constructs: "concepts," "skills," and "problem solving"; these in turn were "defined in terms of student capabilities and the processes needed for successful task performance." The constructs were used to assure a balanced test instrument and were the basis for standards-setting and reporting.

Eva Baker recommended the use of an intermediate strategy to provide a "crosswalk between standards and assessments," but one with the added advantage of providing generalizable assessment models that increase the cost-effectiveness and classroom utility of large-scale assessments. Based in cognitive theory, the CRESST models focus on core types of learning that recur across the curriculum: conceptual understanding, knowledge representation, problem solving, communication, and team work. The models provide specifications for developing assessment tasks and scoring rubrics that are operationalized in specific subject areas and customized to local curriculum emphases and grade levels to be tested. Teachers can use the CRESST models to create classroom instruction and assessment, and to align their practices with established standards and assessments.

Based in cognitive theory, the CRESST models focus on core types of learning that recur across the curriculum...

David Niemi, University of Missouri, and Zenaida Aguirre-Muñoz, CRESST/UCLA, described the application of the CRESST models in Hawaii and elsewhere, where they have been used for assessing history and mathematics. Among the advantages Niemi and Aguirre-Muñoz noted were improved replicability and comparability of tasks and results, enhanced system alignment, greater efficiency, and improved engagement of teachers and the public in all aspects of large-scale assessment.

A number of conference participants stressed that aligning standards and assessments is only one piece of what's required for the success of current reforms. Professional development, curriculum and instruction, incentives and sanctions, resource allocation, district and school infrastructure—all these must be aligned with standards if real progress is to be made.

Aligning standards and assessment with curriculum is essential if student learning is to be affected. But determining such alignment can be complex, as discussed by William Schmidt, Michigan State University. In his research on the Third International Mathematics and Science Study (TIMSS), Schmidt

found that even deciding whether or not a country covered a particular mathematics topic turned out to be anything but simple.

"From a curricular perspective," Schmidt explained, "math is not math everywhere. There is little overall overlap in the countries we studied."

To create the assessment, it was necessary to develop sets of curricularly sensitive items that address areas where groups of countries show overlapping curricula and a sophisticated methodology for characterizing curriculum.

> "...math is not math everywhere. There is little overall overlap in the countries that we studied."

Measuring Progress

THE challenge of accurately assessing student progress was highly salient to a number of conference participants who were struggling with Title I requirements. The key issue was the mandate that schools must show adequate yearly progress sufficient to enable all students to achieve high standards of accomplishment within a reasonable period of time. Like the alignment of standards and assessment, accurately measuring progress is easier said than done.

"From a validation point of view," said Bob Linn, "we have to ask the question: How do we know when we see improvement?"

Linn provided an example from Kentucky, where student achievement as measured by statewide assessments showed significant year-to-year increases while the National Assessment of Educational Progress (NAEP) indicated little or no significant change.

"There may be problems of test comparability from year to year," explained Linn, "caused by differences in conditions of test administration or the degree of alignment between the test and the state's instructional goals." Linn noted other factors that might account for the increases in student achievement on the statewide test including the substantial incentives and sanctions associated with student test performance, possible changes in school populations, students becoming more familiar with the test format, teachers who changed instruction to match the state test content, or some technical design issue associated with performance assessment.

Bengt Muthén, CRESST/UCLA, underscored the difficult analytic challenges of measuring student progress by identifying some key problems including:

- selecting an analysis method that gives the best picture of performance;
- analyzing and interpreting the interaction among multiple growth processes;
- determining the size and duration of instructional and other treatment effects; and
- understanding the aggregate impact of various school experiences and programs and the contributions of individual student characteristics.

Muthén described CRESST's research program to address these problems. Using longitudinal student performance data in a variety of skills areas, Muthén is developing new, multilevel modeling tools to analyze student progress and the contributions of school, classroom, and other factors to such progress. In addition to providing new technical knowledge useful to researchers, this research should provide policy makers with practical guidance for formulating, reporting, and interpreting assessment results within and across schools.

Linking

Linking, a third validity challenge in current assessment systems, was also subject to wide discussion at the conference. States and local districts are developing unique assessments based on their own standards. Yet their publics—parents, community, policy makers, and students—still want to know how their students' performance compares with that of others—from other localities, other states, nationally, and internationally. If students meet local or state standards, does it mean they are nationally or internationally competitive? What are the ground rules for linking results, especially across different types of measures, such as norm-referenced tests and performance assessments?

Conference speakers discussed a number of efforts to link results between the National Assessment of Educational Progress (NAEP), state-by-state NAEP assessments, and international tests such as the Third International Mathematics and Science Study (TIMSS).

For example, Sharif Shakrani, National Center for Education Statistics, presented a number of technical and practical challenges in making links that might permit states to compare their test results nationally and internationally.

"The market-basket approach appears to be one promising option," explained Shakrani, "where states could administer representative samples of items drawn from the full set of

NAEP items along with their state assessments to get good estimates of how students in their state compare with students nationally." Shankrani added that the market-basket approach, applied to NAEP and state assessment, "would have the additional advantage of providing rapid turnaround time, perhaps as quickly as three months, and would allow more frequent NAEP testing in more subjects."

In discussing his agenda for the National Center for Education Statistics, Commissioner Pascal Forgione endorsed research on the market-basket approach and efforts to link state, national, and international data. He also discussed plans to study the feasibility of embedding robust NAEP items in states' non-NAEP assessments to generate more timely and cost-effective state-level NAEP scores. Further, Forgione revealed plans for linking NAEP with other national testing data to build a more comprehensive, cost-effective, and consistent database for policy makers and researchers.

Fairness

While fairness is an integral component of validity, the CRESST model and conference participants identified it as needing concerted attention in current assessment systems. Prompted in part by recent Title I legislation, which will affect approximately 67% of schools in the United States, states and districts are increasingly interested in assessments that will contribute to educational improvements for all students, and they are committed to the testing of all students.

"Results from a recent survey of Council schools...show that 85% of the Council districts have changed or are changing their assessments to align with new national, state, or local content standards."

"The challenge," said Adrienne Bailey, senior consultant for the Council of the Great City Schools (CGCS), "is to get *all* students to perform at higher levels than they have before."

But the urgency to improve schools underscores the need to make sure that the process is fair. "Results from a recent survey of Council schools," said Bailey, "show that 85% of the Council districts have changed or are changing their assessments to align with new national, state, or local content standards." Based on the volume of reform in process, Bailey emphasized that broad community involvement is essential in order to produce standards and assessments that are supported by diverse, urban communities and that help urban students to achieve world-class levels of performance.

"Equity and fairness are no longer simply issues of morality," warned CRESST partner and Yale Professor Emeritus Edmund Gordon; "in educational measurement, they emerge as being at the very core of what our work is about."

Ideally, the quality of inferences from assessments will be judged in terms of their accuracy and appropriateness for all people, of all backgrounds and needs. For this ideal to be approached, every aspect of the assessment process must be fair—from assessment development through administration and interpretation of results. However, although often heralded as a bridge to opportunity, testing and assessment have more often been viewed as unfair to underrepresented minorities and as barriers to educational access.

"To help improve education for all students," added Gordon, "tests must go beyond telling how a student is performing, to giving useful information about how to improve that performance. The system must be committed to adapting to the diverse needs of all students."

Gordon identified four broad categories of fairness issues:

- the political economy of educational assessment;
- limitations in the political and technical capacities of pedagogy and assessment;
- epistemological and theoretical contexts for educational assessment; and
- the technological demands of equitable systems of assessment.

"We must learn to factor into our pedagogical and assessment practices the conditional and situational correlates of human performance," said Gordon. "Why, for example," he asked, "do Black students do better on tests when the test administrator is Black rather than White? To make fair tests we will have to change, expand, and achieve better symmetry among our concepts of knowledge, pedagogy, and intelligence. We will have to honor and accommodate diversity. And we will certainly have to move beyond traditional multiple-choice tests."

"This task," concluded Gordon, "will require a strategic plan for assessment development, and the new CRESST model points us in that direction."

In addition to addressing issues of fairness for students who are currently included in testing, CRESST research is focusing on two groups who traditionally have been excluded from large-scale assessments: students with disabilities and language minority students who are not fully proficient in English.

Students With Disabilities

How many students with disabilities are currently excluded from testing? What are the characteristics of these students? What kinds of accommodations are currently being used? The answers to these key questions are far from clear according to Linda Bond, North Central Regional Educational Laboratory. In a national survey of state-wide testing programs, Bond found tremendous variability in states' estimates of how many students with disabilities were excluded from the state test. In general, anywhere from 5% to 10% of all students were excluded depending on the particular assessment and the state.

In general, anywhere from 5% to 10% of all students were excluded depending on the particular assessment and the state.

Similarly, Daniel Koretz, CRESST/RAND, cited U.S. Department of Education statistics indicating that states' estimates of students with disabilities in their state vary widely, from 5.5% to 15%, with an overall average of about 10%. Estimates of specific disabilities such as mental retardation or learning disabilities are even more variable.

"Differences among states," Koretz suggested, "result mostly from differences in definitions of students with disabilities, rather than real differences in where students with disabilities live."

Problems with counting and defining students with disabilities are not surprising given that states lack clear guidelines defining those students who should and should not be included in their assessments. James Ysseldyke, National Center on Educational Outcomes (NCEO), found that state guidelines on inclusion range from a single descriptive sentence to 60 pages of directions. To help states include more students with disabilities in their testing programs, NCEO is revising and clarifying their guidelines for inclusion practices.

Good descriptive information and sound guidelines, however, may not be enough. Some districts and states are still reluctant to include students with disabilities for fear that low performance will hurt the child's self-esteem and lower overall state scores.

"Nearly 70% of fourth-grade students currently excluded from state NAEP assessments could take those tests," claimed presenter Fran Stancavage, American Institutes for Research. Stancavage urged increased adaptive testing strategies because current assessments are not providing the necessary information about the performance of the least able and the most advanced students. Ysseldyke agreed, observing that "the most difficult measurement issues get addressed at the margins of the

distribution—very high and very low performing students." These factors suggest that assessment developers must make tests simultaneously more challenging and more accessible.

Accommodations will be a key to increasing accessibility, at least for some students. But deciding who needs accommodations, what kinds of accommodations are feasible, and whether accommodations create an unfair advantage for test takers will be a major undertaking. Bond reported that most states have little problem modifying testing conditions for physically disabled students, offering Braille tests for blind students, for example. But accommodations for cognitively disabled students are not as common; and the higher the stakes, the fewer the accommodations because of test validity concerns.

> "The purpose of accommodation from a measurement standpoint is to offset bias in order to make the measurement more valid than it would otherwise be."

"The purpose of accommodation from a measurement standpoint," responded Koretz, "is to offset bias in order to make the measurement more valid than it would otherwise be."

Discussing CRESST's program of research on accommodations and adaptations for special needs students, Koretz described research on the use of paraphrasing for mildly mentally retarded students taking Kentucky's science tests. Issues Koretz is struggling with are the meaning of the scores obtained with and without accommodations and policies for assessing students whose classifications as disabled are ambiguous.

Scott Trimble, Kentucky Department of Education, spelled out his state's strong commitment to include all students in their state assessments. Kentucky districts and schools are advised to use the same accommodations in the assessment that they use in instruction.

"Instructionally relevant accommodations," added Trimble, "such as the use of paraphrasing, extended time frames, and smaller group settings, seem to work fairly well."

But he cautioned that accommodations are not magic solutions to all problems associated with assessing students with disabilities. For example, students may receive special attention in the classroom one year, but not in another year. Trimble also questions whether year-to-year scores for students with disabilities are fair measures of their progress when one year they are tested with accommodations and another year they are not.

Language Minority Students

LIMITED English proficient [LEP] students are more likely to be tested than students with disabilities," said Charlene Rivera, George Washington University. "But LEP students are less likely to be given accommodations," she added. As an example, Rivera reported that in 1994, 17 states required students to pass a high school graduation test to earn a diploma. While 13 of the 17 states permitted accommodations, only 2 states offered tests specifically designed for LEP students. "Scores on tests that assume English proficiency are likely to grossly underestimate LEP students' academic achievement," suggested Rivera.

"For purposes of accountability, improved teaching, and student learning," said Lorrie Shepard, CRESST/University of Colorado at Boulder, "we need assessment systems that can identify student performance on relevant continua of proficiency. For LEP students," added Shepard, "this requires multiple measures that distinguish English language proficiency, native language proficiency, and academic achievement. Such systems are far from being available now. In particular, current measures of English language proficiency are few and limited, yet are essential to understanding LEP students' performance."

"Chicago and the entire state of Illinois are working intensively to develop good measures of English language proficiency," said presenter Carole Perlman, Chicago Public Schools. The state is developing large-scale assessments for LEP students in Grades 3-11 that will be used in 1997. They will include multiple-choice reading tests and writing assessments with both textual and graphic prompts. "In the Chicago Public Schools," added Perlman, "the focus is on giving teachers of bilingual classes the tools to develop and use performance assessments to document both native and English language achievement."

Efforts to develop appropriate accommodations for LEP students are just beginning. Translating tests into students' native language seems like a straightforward solution; and indeed, the two states Rivera identified as offering alternatives to their high school graduation tests use translations. But if students haven't learned the content in their native language, the value of a translated test is debatable. "Furthermore," Shepard warned, "linguistic and cultural differences make exact translation impossible, casting doubt on the equivalency of scores such tests yield." She called for more small-scale, focused research to provide good information for making accommodation decisions. Agreeing that translations do not meet the needs of all LEP students, Rivera nonethe-

> . . . if students haven't learned the content in their native language, the value of a translated test is debatable.

less called for their increased use, at least in the short run, "especially for students who enter schools with a high degree of literacy in their native language."

Simplifying test language is another accommodation option. Based on his analysis of NAEP data, Jamal Abedi, CRESST/UCLA, cited evidence that complexly worded mathematics questions depress LEP students' scores. Abedi discussed CRESST research indicating that simplified wording significantly improved scores, at least for some students. Additional projects are underway to examine the effects of linguistic and other adaptations on limited English proficient and fully English proficient students of varying abilities.

Using mixed-ability, collaborative work groups also may be an effective—and instructionally relevant—form of accommodation. During her research in classrooms with high LEP populations, presenter Noreen Webb, CRESST/UCLA, found that low-ability students scored better on performance assessments when they worked on similar tasks in groups with high-ability students prior to the test. "High-ability students' scores were not affected," said Webb.

Although there are hints of promising directions, it is clear that there is much work to be done to develop valid, fair, and useful tests for all students. A complicating factor is that the work must be done with an eye toward making assessments that the public understands and trusts.

Credibility

"The most valid and fair assessment systems imaginable will fail if they lack public credibility," said Eva Baker during her conference presentation. "Validity without credibility produces assessments that have no life span and whose findings are contended, diminished, or dismissed," added Baker.

Several presenters addressed the importance of CRESST's third prerequisite for utility, arguing that public communication and engagement are essential in establishing assessment system credibility.

Secrecy, driven partly by the legitimate need for test security, has long been a trademark of the measurement community. But the public is increasingly reluctant to accept assessments—new or otherwise—on blind faith. As a result, many members of the assessment community have found that they need improved communication and public relations skills to complement their technical skills.

Lorraine McDonnell, CRESST/University of California, Santa Barbara, addressed the broader social and political context, which demands better communication of assessment information. Citing recent polls, McDonnell noted that only 25% of the public trusts government institutions to do the right thing all or most of the time, and only 25% of the voters, who decide whether or not to support public education with taxes, even have children in

school. Consequently, new assessment systems must survive in a context of mistrust and limited public understanding of education. "In this environment," McDonnell noted, "curricular standards and assessments become the focal point for many contested social values, not just about what is important to learn, but about how we define the good society and how those ideals should be passed on to succesive generations." Based on her research on the politics of education reform in several states, McDonnell suggested several guidelines for making assessments politically and publicly credible.

> "...leadership has to come, at least partially, from people who are electorally accountable..."

"First, where political will is lacking to make a needed long-term investment," said McDonnell, "an incremental approach may actually yield better results than a comprehensive approach."

"Second, if a state decides to engage in substantial reform," McDonnell argued, "strong political leadership is necessary. That leadership has to come, at least partially, from people who are electorally accountable," added McDonnell, "not just from the education establishment and non-elected officials. Elected officials," she pointed out, "are in regular contact with constituents, and the constraint of their two- or four-year electoral cycle helps them bring a valuable, real-world perspective to the process."

"Third," McDonnell asserted, "the development of new curriculum standards and assessments cannot solely be a technical process with participation limited to experts. ... Public participation in open, two-way dialogues is very important," McDonnell noted, "because it involves public deliberation about what skills and knowledge are most important for a productive life and active citizenship." Acknowledging that building public consensus is a very difficult process, McDonnell warned that to avoid it would be to make a mockery of the notion of common standards.

"Communicating a topic as complex as assessment," said Leah Lievrouw, CRESST/UCLA, "is a formidable challenge in any modern, diverse society. ... Despite improvements in mass communication, we live in an era of separation characterized by high levels of intra-group conversation and low levels of inter-group communication," she added. Consequently, we may tend to target large media outlets while ignoring the types of smaller market electronic and print media tailored to many linguistic and ethnic minorities. "As a result," explained Lievrouw, "our message may not get across to many of these important groups, and we need to rethink our media strategies."

Richard Colvin, *Los Angeles Times*, emphasized the crucial value in building public credibility, something that was not done by the California State Department of Education during the California Learning Assessment System crisis. "CLAS was a forty-million-dollar mistake," said Colvin, "not because it produced invalid or unfair results, but because of widespread public distrust caused by perceptions that there were weird things on the test." Citing the bunker mentality that led to the demise of CLAS, Colvin urged researchers and test makers to engage in open and understandable communication with the public. He offered several tests that he feels assessments must pass to achieve public credibility:

- *The "barber chair" test.* Ordinary people should be able to discuss the assessment in ordinary social situations. Assess familiar content that the public thinks children need to get along in the world.

- *The "realtor" test.* Test scores affect housing values and, consequently, influence homeowners' support for local schools. Report scores in a format simple enough that realtors can use them as a closer.

- *The "newspaper" test.* Newspapers have limited space for even the most important stories. Give us results that we can fit into two or three columns.

Challenging the research community to make testing understandable to the general public, Colvin advised that "there is a pace you can walk at that the public can follow, or you can run out ahead and lose everybody."

McDonnell's, Lievrouw's, and, particularly, Colvin's remarks challenged—even nettled—some of the conference attendees, but the themes were strongly endorsed by other CRESST presenters who have been working hard to build public support for high standards and improved assessments. Moreover, these presenters universally argued that credibility and successful implementation was not possible without active, broad-based, public participation.

In a roundtable presentation, four state education officials—Duncan MacQuarrie, Washington; Wayne Martin, Colorado; Doris Redfield, Virginia; and Catherine Smith, Michigan—identified key constituencies that must be formally included in the process for any state-level reform to be successful. The list includes teachers, students, parents, school administrators, school board members, higher education officials and admissions officers, education researchers, state legislators, the governor, representatives of the business community, and the media. Presenters noted that teachers are almost always deeply involved in the process from the beginning, but that other groups, particularly school administrators and parents, should be involved more directly and earlier than usual.

Echoing Colvin's advice, these presenters also stressed the importance of making certain that standards and assessments include items that represent the public's general understanding of a subject and what's important, for example, a computation standard in mathematics or a grammar/spelling standard in writing.

Discussing district-level reform, Los Angeles Board of Public Education President Mark Slavkin identified several keys in building and maintaining credibility at the district level.

Slavkin identified jargon as a major threat to credibility...

"To keep public and political credibility," said Slavkin, "it is necessary to keep one foot in the old [norm-referenced assessments] as we move to the new standards-based performance assessments."

Slavkin identified jargon as a major threat to credibility, recommending that everything be publicly disclosed in the process of development, and emphasizing the importance of keeping the media updated about progress along the way.

But foremost, in Slavkin's opinion, is "buy-in" from the very beginning by parents, teachers, and community members. One example of a successful, if not always smooth, effort to get such commitment has been a three-year project to develop language arts standards, curriculum, and performance assessments in the Los Angeles Unified School District. Described by Charlotte Higuchi, a CRESST partner and LAUSD teacher, this project involved a large and diverse group of teachers, parents, and community members whose input shaped the reform from the beginning.

Los Angeles Unified School District Superintendent Sidney Thompson also emphasized the importance of buy-in at the school level. "The people who have to do it at the school site have to own it; they have to believe it can be done, and if they believe that and bring the parents into it, then we have a pathway to get us there."

UTILITY

VALIDITY, fairness, and credibility are necessary, but not sufficient, for system utility. Because assessment systems usually serve multiple purposes and users, often with different and competing needs, it becomes very difficult to design a system useful for all purposes and people. Randy Bennett, Educational Testing Service, expressed concern that the utility of a complex, multipurpose assessment system would be similar to the Swiss Army Knife, serviceable in a pinch for all sorts of jobs from removing screws to opening cans, but not ideal for any one of them.

Like a sensitive ecosystem, the utility of an assessment system will likely be greatest when the individual components are harmonious with the processes that link them. If any process or component is disrupted, the entire system suffers. Unfortunately, the political environment in which assessment exists is volatile and urgent as Judith Billings, then superintendent of education for the state of Washington, noted.

> "...if you don't give us good information about what works, and soon, those of us responsible for implementing assessment reforms may well perish."

"You [academicians] may publish," said Billings, "whether new assessments work or not. But if you don't give us good information about what works, and soon, those of us responsible for implementing assessment reforms may well perish."

In Washington, as in virtually all states, one of the purposes of assessment reform has been to change teaching and, thus, to improve student learning. Many conference presenters this year focused on an element central to the CRESST assessment model, teacher capacity building and teachers' changing instructional practices as a result of changes in assessment methods.

Issues in Improving Instruction and Teacher Capacity Building

Teachers are a key to the credibility of assessment reform and essential to its success," said presenter Sid Thompson. But Marilyn Monahan, secretary-treasurer of the National Education Association, warned against the assumption that teachers will be able to immediately embrace reform. Standards-based assessment reform demands change in instructional practices. Monahan stated that teachers welcome this, but that those who want reform must invest in teacher knowledge through professional development.

"The journey from what takes place at this conference to what takes place in teachers' classrooms is long, complex, and unpredictable," cautioned Monahan.

Agreeing with Monahan, CRESST partner Hilda Borko found in her research with teachers and students that teachers were interested in assessment and instructional reform, but felt they didn't have the time or expertise to develop alternative assessments on their own. Borko found that teachers initially inserted performance assessment into their instruction; that is, the new assessments were added into an already busy instructional program. While some teachers began to incorporate these assessments into their ongoing instruction by the end of the first year of the reform effort, it often was not until the third year of the project, with

help from Borko and others, that teachers were able to integrate assessment and instructional reforms into their daily classroom activities.

Maryl Gearhart and Megan Franke, CRESST/UCLA, reported on action research projects focusing on assessment reform in mathematics. The integration of assessment into instruction, they argued, is essential to the realization of mathematics reform envisioned by the National Council of Teachers of Mathematics standards. Gearhart and Franke's research showed that teachers need deeper understanding of mathematics and children's mathematical reasoning in order to implement new pedagogies at more than a superficial level. For example, although many teachers participating in their projects began to ask children to share their thinking, only some teachers were able to probe with specific and substantive questions or to guide analytical discussions of student problem-solving strategies.

> The journey from what takes place at this conference to what takes place in teachers' classrooms is long, complex, and unpredictable...

"Classroom instruction and performance assessment should be inseparable aspects of the education experience for language minority students," urged Richard Durán, CRESST/University of California, Santa Barbara. He recommended that teachers gather multiple forms of evidence of student performance—classroom tests, graded projects, student self-assessments, and videotapes revealing students' fluency with learning tools.

Agreeing with Durán, Thomas Romberg, University of Wisconsin, Madison, argued that "teachers need multiple assessment strategies to accurately measure student performance," adding that "in mathematics, teachers need to know if students can add, subtract, multiply and divide, and if they can put these piecemeal skills together to solve routine and nonroutine problems."

"But teachers also need to know if students can explain *why* an answer is correct," added Romberg, "and how students navigate the problem-solving process." He suggested creating situations that enable teachers to gather information from listening to students' explanations, observing students at work, and examining the products of their work. Large-scale assessment can signal students' level of performance, but teachers need continuous, detailed evidence of students' learning processes and understanding to fully support the multiple assessment process.

That large-scale, standards-based performance assessments are not typically designed to meet the information needs of the classroom teacher was also noted by Phil Daro, director for assessment development for the New Standards Project. Daro recommended

that one method to help teachers understand what standards demand of their students is to make performance assessments look like well-constructed teacher-tests.

"Teachers do standards-based instruction and performance testing all the time," said Daro. "The trick is to clarify the standards and assessments sufficiently for teachers to recognize the parallels with their own practice," added Daro, "but not so much that the true complexity of the reform effort is lost."

"We need to use large-scale performance assessment systems to define standards and focus school-level efforts on student work linked to standards," agreed Lynn Winters, Long Beach (CA) Unified School District. She emphasized the need to engage teachers in continuing professional activities to enable them to implement effective, ongoing, standards-based classroom instruction and assessment.

"Professional development is the hardest sell in the reform marketplace," admitted Catherine Smith, Michigan State Department of Education, but she added that it was a vital component to the success of any reform effort.

USING TECHNOLOGY TO CREATE NEW POSSIBILITIES

At least a few of the major problems presented by complex assessment systems might be resolved by improvements in technology according to several conference presenters. Clearly, technology can help with managing, linking, and disseminating assessment results. But technology also may permit test developers to increase authenticity and open up the range of modalities and systems of representation used for assessment.

"Future generations of tests will need to tap nontraditional constructs, base test designs on cognitive principles, and increase the diversity of problems types," noted Randy Bennett, Educational Testing Service. In spite of current logistic problems, Bennett predicted that large-scale assessments would soon include computer-based presentations of problem types not possible with paper-and-pencil tests. Bennett shared multimedia prototype items using historical speeches and newscasts to illustrate the potential of presenting and asking students to respond to "dynamic stimuli."

Ron Stevens, CRESST/UCLA, demonstrated the use of neural network technology to permit real-time assessment of complex problem solving. In one of Stevens' prototypes, medical students were presented with realistically sketchy information about a patient's symptoms, a set of diagnostic tests that they could order, and a "library" of reference materials. As the students worked through the options presented by the computer program, their choices were recorded and could be compared with patterns of hypotheses generated by expert diagnosticians investigating the same problem.

Based on lessons learned from his efforts to develop computer-based assessments of group and teamwork processes, Harold O'Neil, CRESST/University of Southern California, noted a number of problems endemic to technology projects. These include the costs and time for software design and development, inadequacies of existing telecommunications technology, unavailability of sophisticated technologies in public schools, and the complexity and cost of maintaining test security. O'Neil also shared the substantial progress his group has made in using technology to measure the quality and quantity of individual contributions and group problem solving. He pointed particularly to the potential of collaborative concept mapping where students work in teams through networked technology to create and revise concept maps. Collaborative concept mapping makes possible the real-time assessment and reporting of deep understanding and teamwork performance—an example of a potentially useful and cost-effective near-term application of technology.

CONCLUSION

In their closing remarks to the 1996 CRESST conference, Eva Baker and Robert Linn acknowledged the formidable challenges ahead. Among them are assuring system validity; supporting the alignment between standards and assessments; promoting fairness; addressing the technical challenges of measuring progress and linking different assessments to address the needs and purposes of many audiences at multiple levels; improving schools' and teachers' capacity; and productively engaging the public. "These are all priorities for the research community," said Baker and Linn. "They will call on the best of our technical skills along with very considerable sociopolitical prowess."

"I think a key lesson of the past two days," concluded Baker, "is that we must approach the assessment challenges we've discussed through greater collaboration. It's clear that we share a commitment to improve education. Let's move forward together."

1997 CRESST Conference

September 4 through September 5, 1997

Sunset Village Conference Center
UCLA Campus

Registration materials and complete details will be available in the Summer 1997 *CRESST Line* and on the CRESST World Wide Web site, www.cse.ucla.edu. Anyone requiring early registration may contact Mary Wilby, CRESST/UCLA, 10920 Wilshire Blvd., Ste. 900, Los Angeles, CA 90024; e-mail: mary@cse.ucla.edu; phone: (310) 206-1532.

CRESST Reports Available Now

The following CRESST reports are available by calling Kim Hurst, (310) 206-1532, or sending a message to Kim at: kim@cse.ucla.edu.

Reforming Schools by Reforming Assessment: Consequences of the Arizona Student Assessment Program (ASAP): Equity and Teacher Capacity Building
Mary Lee Smith
CSE Technical Report 425, 1997 ($9.00)

In this study, Mary Lee Smith and other researchers focused on how schools changed as a result of state-mandated standards and assessments.

The Politics of State Testing: Implementing New Student Assessments
Lorraine McDonnell
CSE Technical Report 424, 1997 ($5.00)

Lorraine McDonnell continues her syntheses of innovative state assessment programs in Kentucky (Kentucky Instructional Results Information System), California (California Learning Assessment System), and North Carolina.

Teachers' Developing Ideas and Practice About Mathematics Performance Assessment: Successes, Stumbling Blocks, and Implications for Professional Development
Hilda Borko, Vicky Mayfield, Scott Marion, Roberta Flexer, and Kate Cumbo
CSE Technical Report 423, 1997 ($3.00)

This study focuses on the change process experienced by a group of third-grade teachers as they implemented mathematics performance assessments in their classrooms. Based on workshop conversations and interviews between teachers and the research/staff development team throughout a single school year, the team reached five major conclusions.

New Writing Assessments: The Challenge of Changing Teachers' Beliefs About Students as Writers
Shelby Wolf and Maryl Gearhart
CSE Technical Report 422, 1997 ($3.00)

During a two-year collaboration with elementary school teachers, Wolf and Gearhart examined ways that teachers' beliefs about their students as writers mediated their investment in new methods of assessing students' writing.

Teachers' Beliefs About Assessment and Instruction in Literacy
Carribeth Bliem and Kathryn Davinroy
CSE Technical Report 421, 1997 ($3.00)

Bliem and Davinroy further investigate teachers' beliefs about assessment and its connection to instruction in literacy. Most of the data were drawn from transcripts of biweekly meetings between the research team and third-grade teachers using performance assessments in their classrooms.

The Politics of Assessment: A View From the Political Culture of Arizona
Mary Lee Smith
CSE Technical Report 420, 1996 ($3.00)

Mary Lee Smith traces the events of the Arizona Student Assessment Program (ASAP), an innovative multiple assessment program that grew out of discontent with mandated standardized testing in Arizona.

Implications of the OECD Comparative Study of Performance Standards for Educational Reform in the United States
Eva L. Baker
CSE Technical Report 419, 1996 ($3.00)

In this report, Eva Baker explores the implications for education reform in the United States of an OECD study of performance standards. Using a general model of educational reform, Baker analyzes the meaning of performance standards in the United States, addressing key influences of tradition, diversity, control, and participation.

Assessment and Instruction in the Science Classroom
Gail Baxter, Anastasia Elder, and Robert Glaser
CSE Technical Report 418, 1996 ($2.50)

Findings from this study of fifth-grade students provided further evidence that critical differences exist between students who think and reason well with their knowledge and those who do not.

CSE Technical Reports may also be found on the CRESST Web site: www.cse.ucla.edu.

Fold here and mail.

Place postage here.

UCLA Center for the Study of Evaluation
10920 WILSHIRE BLVD SUITE 900
LOS ANGELES CA 90024-6511

Products Order Form

CSE Number	Title	No. of Copies	Price Each	Total

Postage and Handling

Subtotal of: Up to $10 add $1.50
$10.01 - $20 add $2.50
$20.01 - $50 add $3.50
over $50 - add 10% of Subtotal

(Please prepay orders less than $10.00.)

Order Subtotal	
Postage & Handling	
California Residents add 8.25%	
TOTAL	

Name or Mailing Address PRINT or TYPE

☐ I would like to receive free copies of *The CRESST Line* and *Evaluation Comment*.

UCLA Center for the Study of Evaluation
1320 Moore Hall/Mailbox 951522
Los Angeles CA 90095-1522
www.cse.ucla.edu
Address Correction Requested (ED 4M)

NON PROFIT ORG.
U.S. POSTAGE
PAID
U.C.L.A.

Dick Hays
Deputy Asst. Secretary
US Dept. of Education
555 New Jersey Ave., NW
Washington DC 20208